Science gateways, hybrid clouds, and distributed workflows for big data (and other data too)

Track: Analytics

Dr. Craig A. Stewart

stewart@iu.edu ORCID ID 0000-0003-2423-9019
pti.iu.edu

Executive Director

PERVASIVE TECHNOLOGY INSTITUTE
INDIANA UNIVERSITY

License terms and citation

- Please cite as: Stewart, C.A. 2018. Science gateways, hybrid clouds, V4I, and distributed workflows for big data. Presented at Indy Big Data Conference. https://www.indybigdata.com. 26 September 2018. Indianapolis, IN. Available from http://hdl.handle.net/2022/22932.
-

The Indiana University Pervasive Technology Institute (IUPTI) transforms new innovations in cyberinfrastructure and computer science into robust tools and supports the use of such tools in academic and private sector research and development. IUPTI does this while bolstering the Indiana Economy and building Indiana's 21st century workforce.

About the Indiana University Pervasive Technology Institute (@IU_PTI on twitter)

- IU_PTI is Indiana University's initiative for advanced information technology research, development, and delivery in support of scientific discovery, scholarly investigation, and artistic creation.
- Information technology today pervades scholarly discovery in the humanities, research in all areas of the sciences, and the processes of artistic creation. The "Pervasive" in the name IU Pervasive Technology Institute reflects the foundational importance of computer science, informatics, cyberinfrastructure, and information technology research to most of what is done in academia and industry today.

IU_PTI Is a collaborative organization with seven affiliated centers:

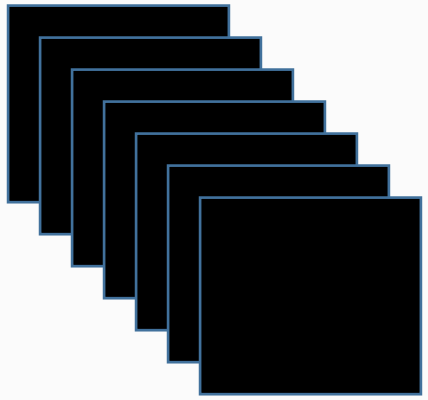
WASHINGTON BIG A TECHNOLOGY CONFERENCE

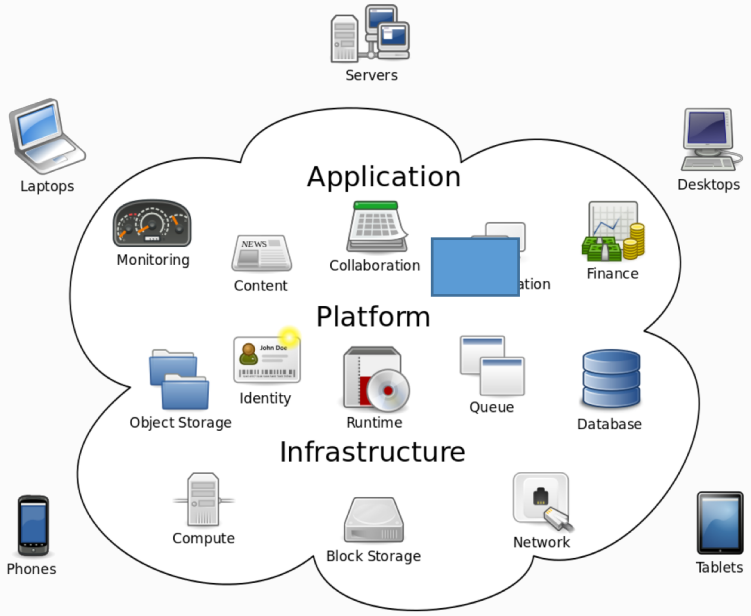
- Center for Applied Cybersecurity Research, led by Von Welch. https://cacr.iu.edu/
- Center for Science Gateways and Distributed Cyberinfrastructure Research, led by Dr. Marlon Pierce. https://sgrc.iu.edu/
- Data to Insight Center, led by interim director Dr. Inna Kouper. https://pti.iu.edu/centers/d2i/index.html
- Digital Science Center, led by Distinguished Professor Geoffrey C. Fox. https://www.dsc.soic.indiana.edu
- Hathi Trust Research Center, led by Professor John Walsh. No active website yet.
- National Center for Genome Analysis Support, led by Dr. Thomas G. Doak. https://ncgas.org
- Research Technologies, led by Associate Vice President Matthew R. Link. https://pti.iu.edu/centers/rt/

IU_PTI Is a collaborative organization with seven affiliated centers:

- Is a collaboration among multiple subunits of IU:
  - Office of the Vice President for Information Technology
  - University Information Technology Services
  - School of Informatics, Computing, and Engineering
  - College of Arts and Sciences
  - Maurer School of Law
  - Kelley School of Business
- Is constituted flexibly to bring together the intellectual and organizational assets of IU to important problems facing society today, while making the organizational structure of IU irrelevant to the collaborations in which it is engaged

Life is complicated. Or at least computing is.

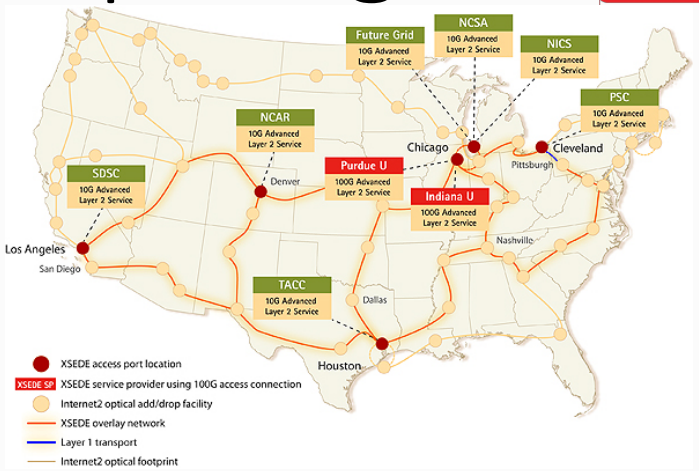




Cloud computing

(The things above are shown as black Boxes for a reason)

And HOW many vendors?



XSEDE – xsede.org

Open Science Grid – opensciencegrid.org

First a moment: why does this all matter anyway?

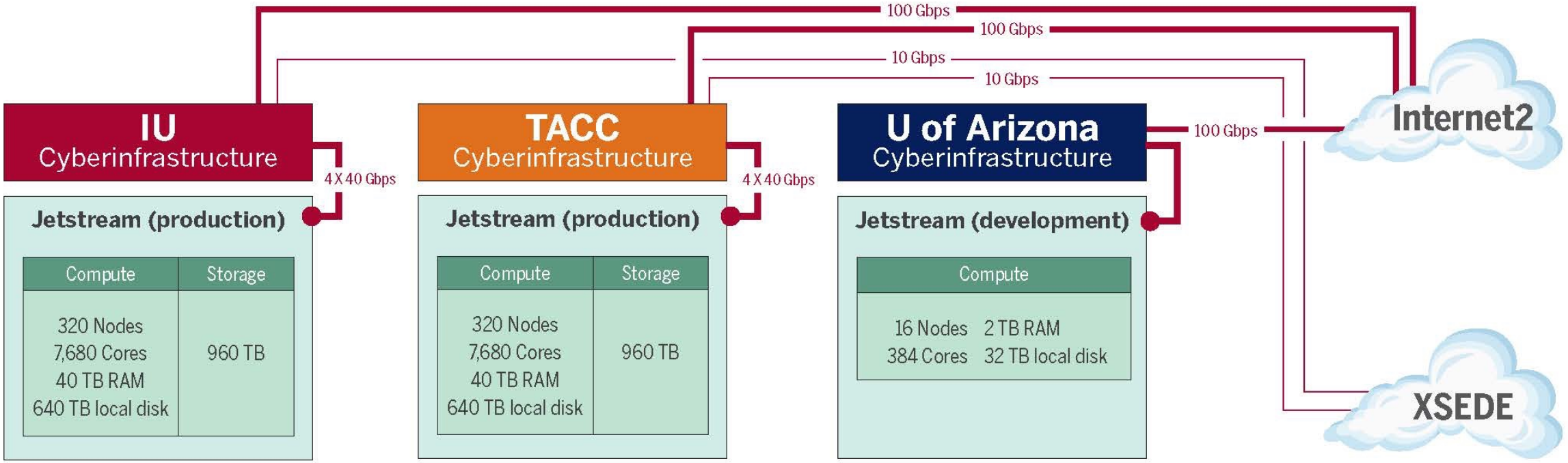
- When you have engineers designing critical turbine components
- Doctors trying to find new treatments for specific types of cancers
- Economists and social scientists trying to identify food deserts
- Students learning how to do research
- …..

- Do you want them to focus on learning how to use a bunch of arcane computing systems that are going to be outmoded in 2-3 years anyway, or do you want them doing their engineering work, their research, or their learning activities?
- At the end of the day it's about making people's lives better (and longer)

In the rest of this talk

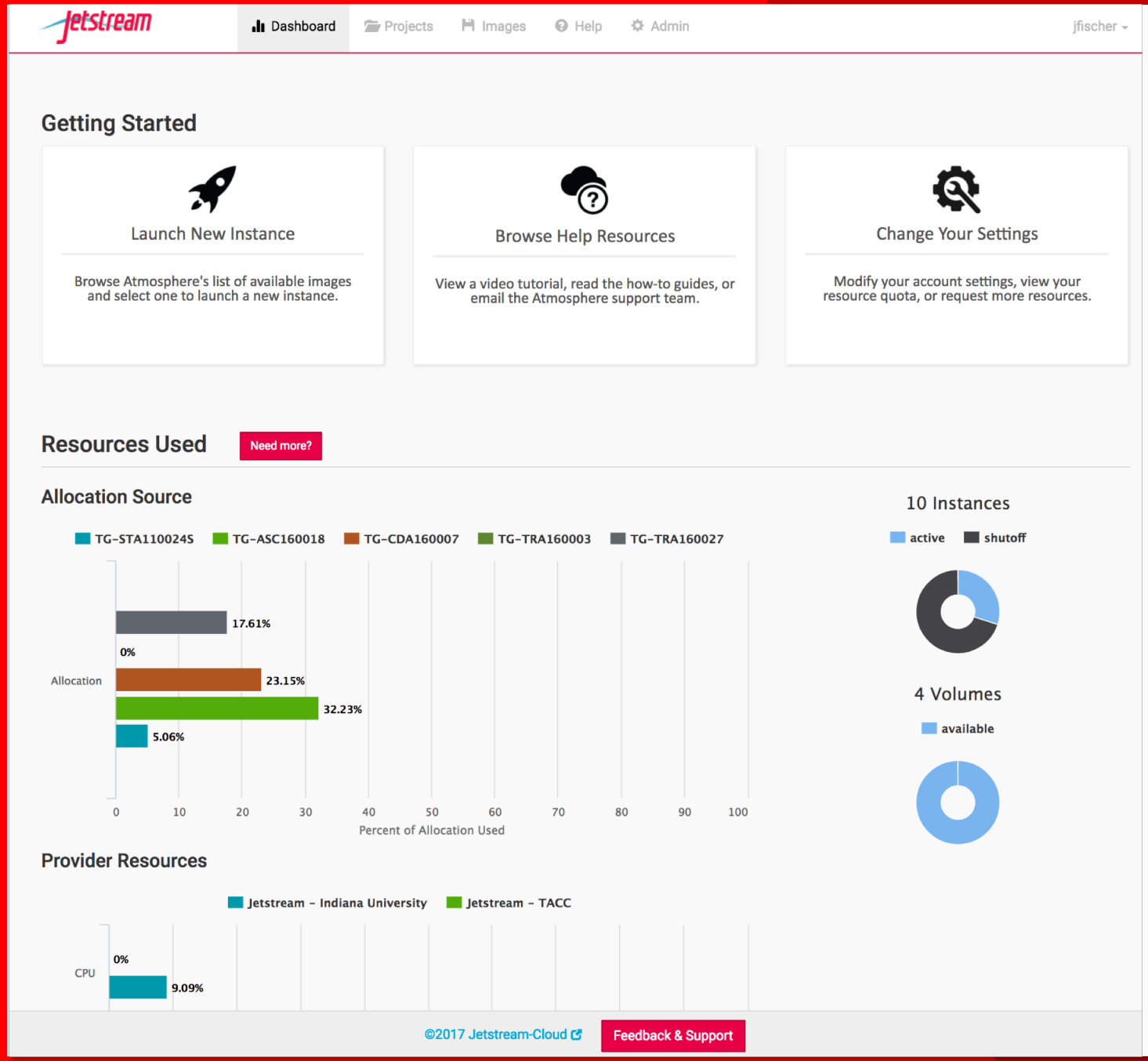
- Jetstream as an example of a particularly useful and popular cloud system
- Science Gateways as an approach to heterogeneous cloud and high performance computing system integration
- V4I as an example of the utility of private-public partnerships in approaching big data

Jetstream System Overview



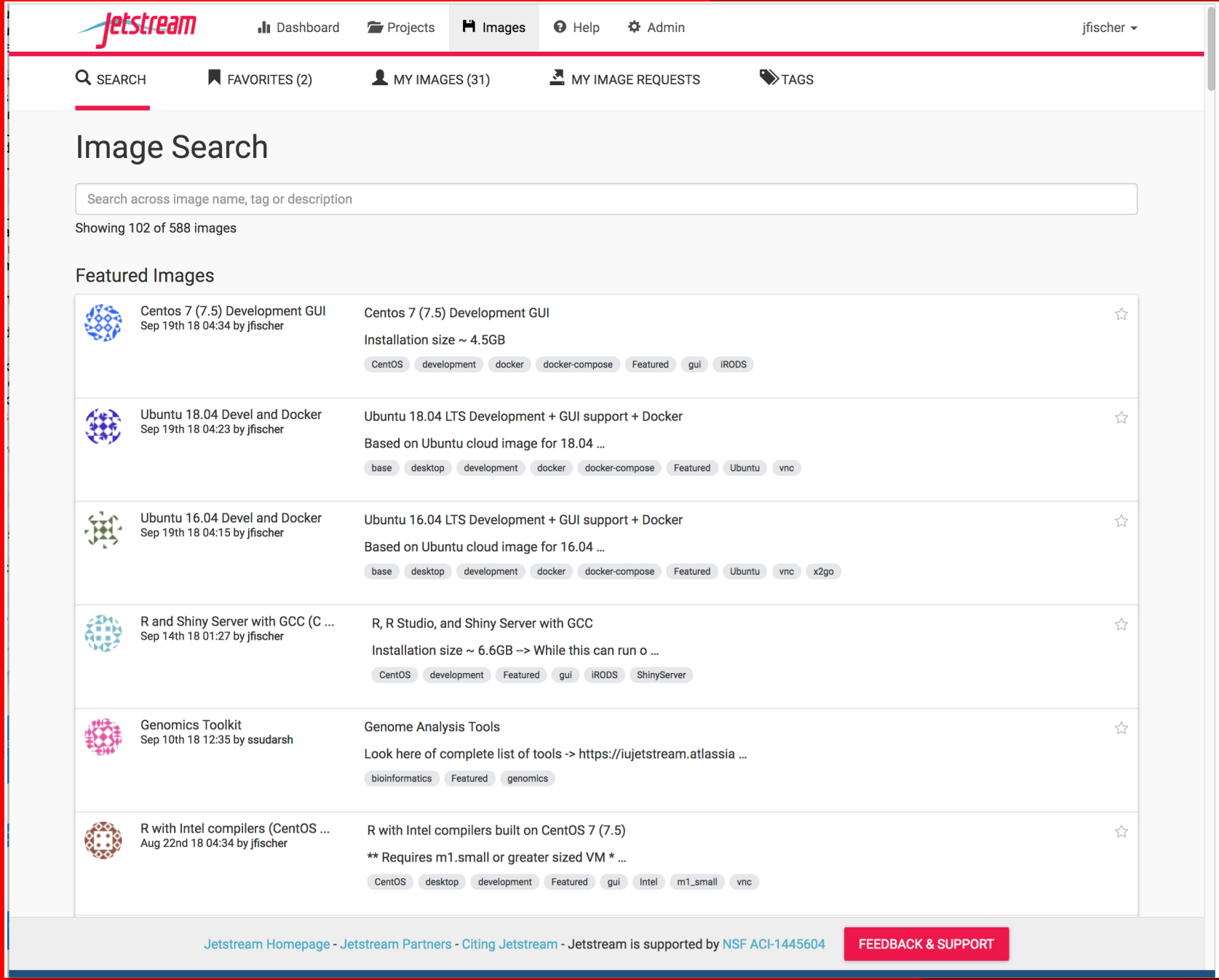
Jetstream
http://jetstream-cloud.org/

funded by the National Science Foundation
Award #ACI-1445604



Jetstream atmosphere web interface

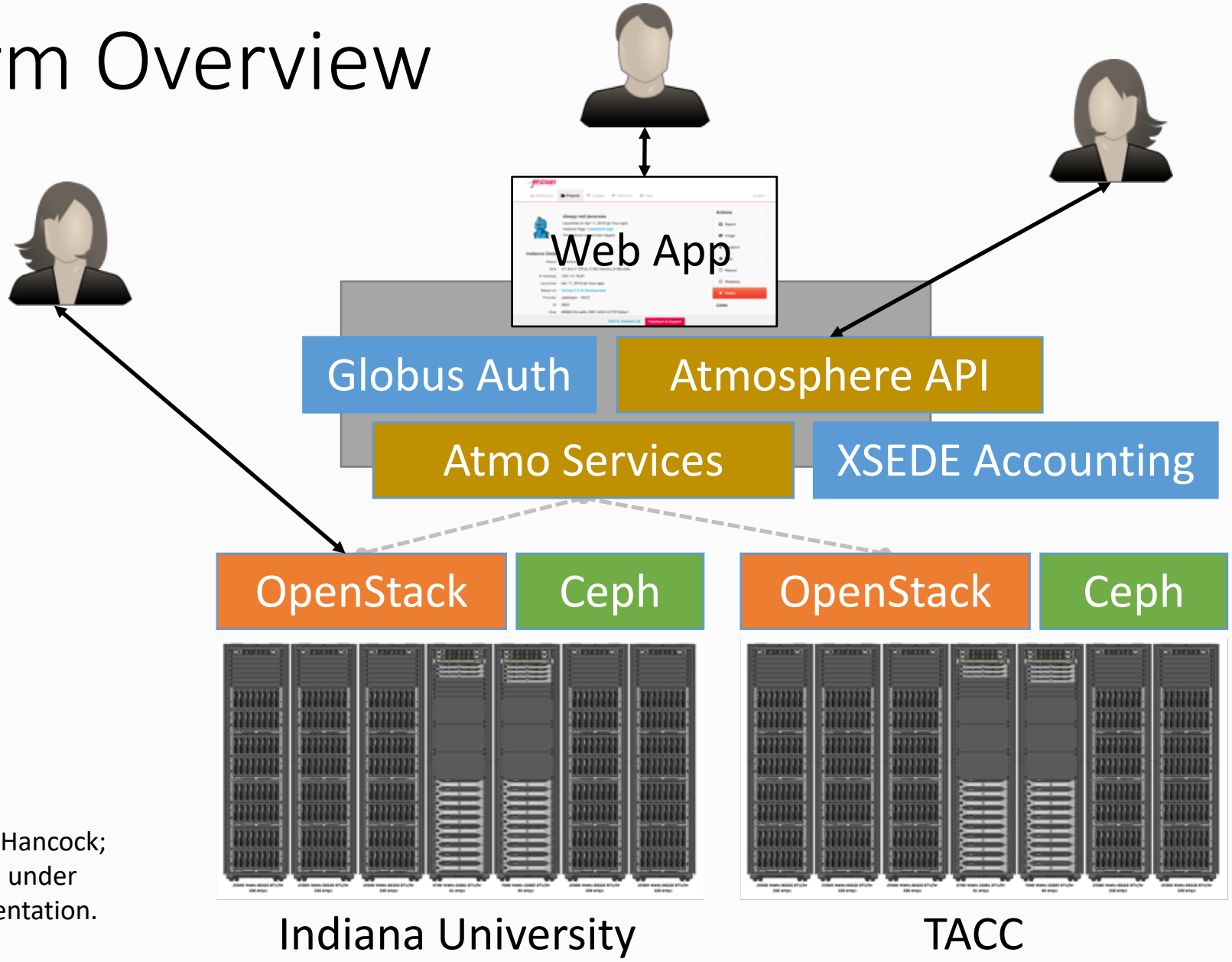
This slide courtesy Jeremy Fischer; jeremy@iu.edu; released under default license for this presentation.



WASHINGTON BIG
A TECHNOLOGY CONFERENCE

This slide courtesy Jeremy Fischer; jeremy@iu.edu; released under default license for this presentation.

Platform Overview



This slide courtesy David Y. Hancock; dyhancoc@iu.edu; released under default license for this presentation.

Expanding the reach: Jetstream REU Program

- NSF Supplement for undergraduates
- 4 students participated in 2017
- 6 students participated in 2018

- REU student videos on YouTube https://www.youtube.com/user/I UPTI

Jetstream
http://jetstream-cloud.org/

funded by the National Science Foundation
Award #ACI-1445604

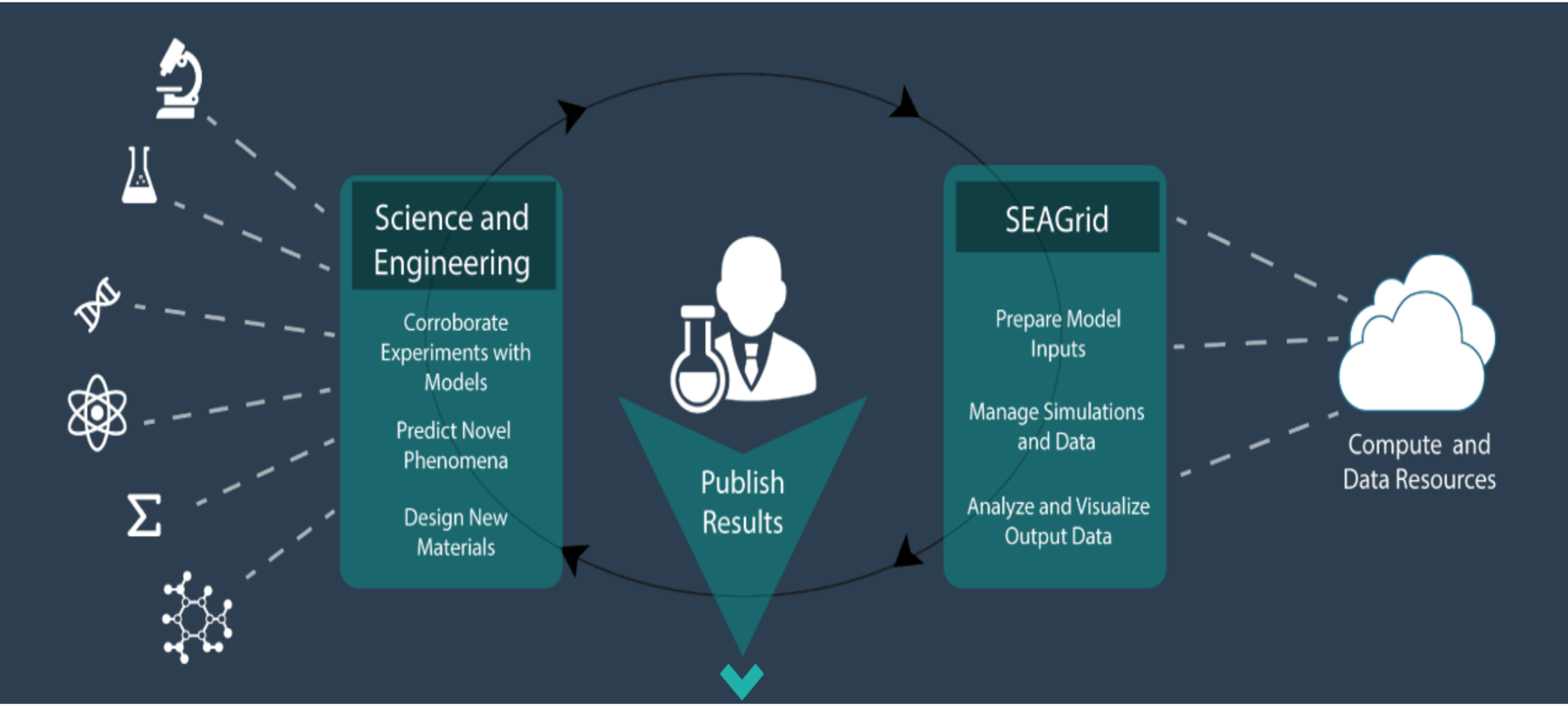
This slide courtesy David Y. Hancock; dyhancoc@iu.edu;

Discipline or area of interest	#of Jetstream allocations	SUs allocated on Jetstream	% of SUs allocated on Jetstream	% of all SUs allocated on other XSEDE-supported systems
Astronomy	2	1,108,096	3.04%	8.61%
Atmospheric Sciences	4	2,752,400	7.55%	3.73%
Biological Sciences	57	5,199,000	14.27%	4.95%
Campus/Domain Champions	123	6,105,500	16.76%	0.09%
Computational Science	11	1,150,000	3.16%	0.92%
Computer Science	15	4,944,302	13.57%	1.8%
Education Allocations	24	2,847,600	7.82%	0.01%
Engineering	1	100,000	0.27%	3.81%
Geosciences	10	1,978,400	5.43%	2.87%
Humanities/Social Sciences	10	560,000	1.54%	0.45%
Molecular Biosciences	8	4,647,520	12.75%	17.65%
Network Science	3	200,000	0.55%	0.06%
Ocean Science	3	230,000	0.63%	1.30%
Physics	4	2,252,400	6.18%	16.43%
Training & Development	11	2,362,000	6.48%	0.16%

What is a Science Gateway?

- Science Gateways are web interfaces and middleware for integrating distributed computing and data, automating expertise, controlling access, managing results, and speeding up your critical computational workflows
- Science gateways encode expertise
  - Running specific scientific applications and workflows
  - Running jobs on diverse, local and nonlocal machines
  - Moving data to and from world-wide resources
- Science gateways enable sharing of results
- Science gateways make results **recoverable, reproducible and reusable**
- **Science gateways form a means of access to heterogeneous cloud services and provide "future-preparing" in face of constantly changing cloud technologies**

Free American Chemical Society (ACS) Workshop on Science Gateways, New Orleans, LA, Mar 20, 2018



SeaGrid Capabilities

Molecular Dynamics (MD)

Time evolution of large material and biological systems to predict dynamic structural and energetic characteristics using Applications such as Lammps, Amber, NAMD using mostly empirical force fields including reaction force fields.

Computational Chemistry

Optimize and characterize molecular and periodic structures and predict thermodynamics and kinetics using computational chemistry applications using Ab initio, Semi-empirical, Force-field based codes such as Gaussian, Gamess, Tinker, DFTB+.

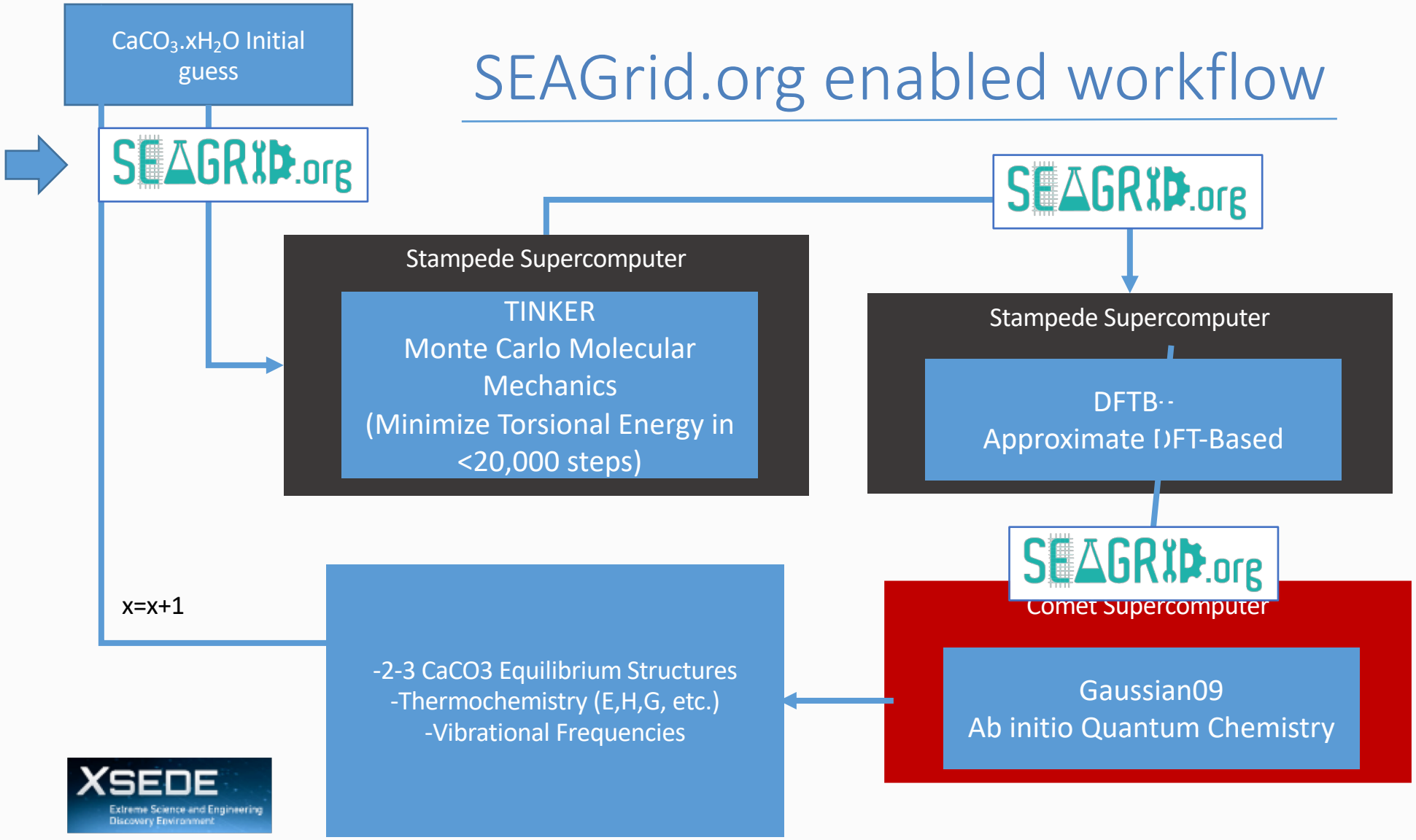
Structural Mechanics

Finite element analysis of engineering structures and components for modeling static and low-speed dynamic events both in the time and frequency domain using applications such as Abaqus.

Fluid Dynamics

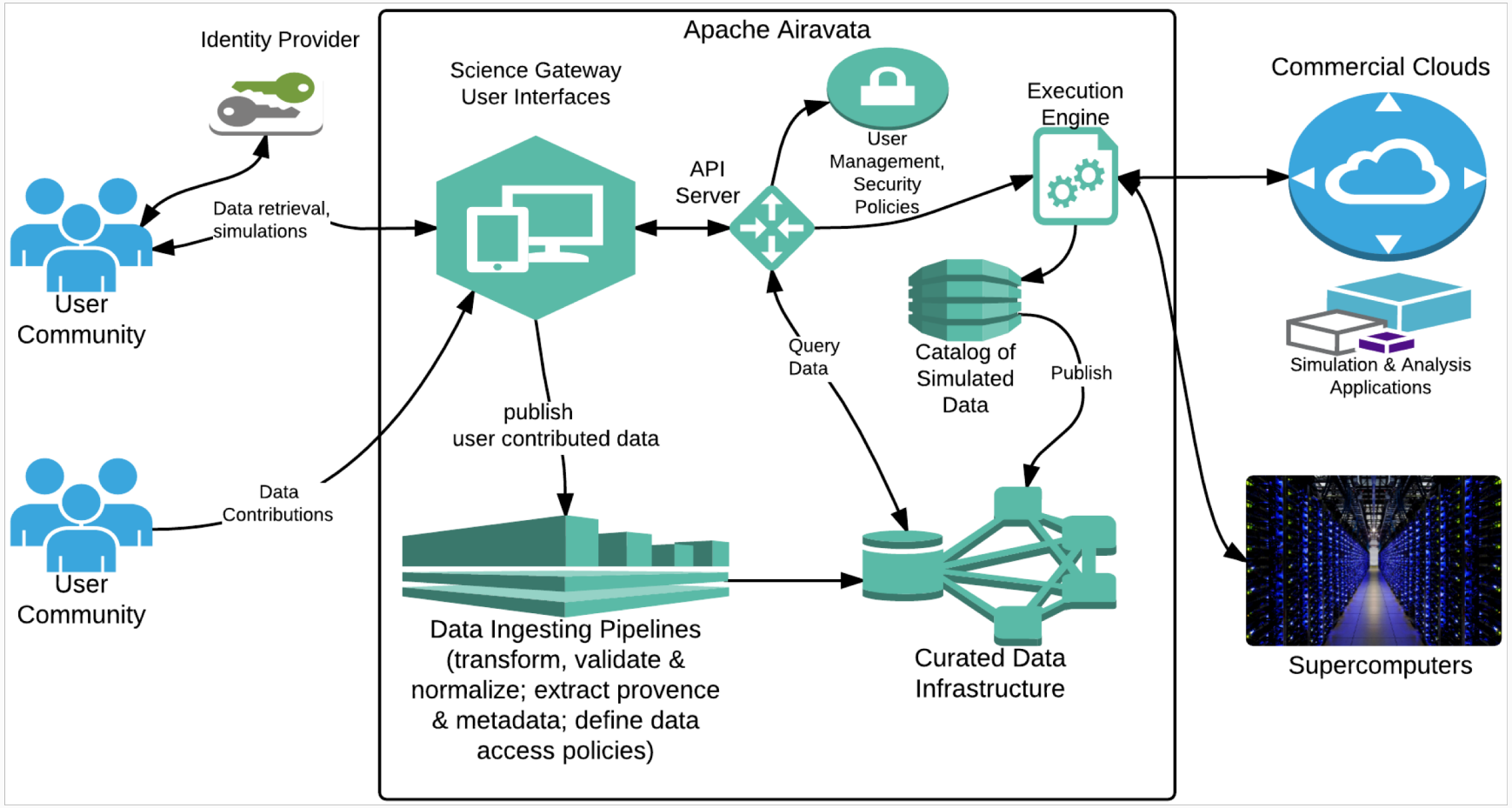
Modeling flow of gases and liquids under various conditions using applications such as Nek5000 and OpenFOAM.

SEAGrid.org enabled workflow



Lopez-Berganza, et al. *J Phys. Chem. A*(2015)

Science Gateway Architecture in General



This slide courtesy Sudhakar Pamidighantam, pamidigs@iu.edu; released under default license for this presentation.

Apache Airavata

- Apache Airavata is software for building science gateways.
  - Don’t start from scratch
- Airavata-based gateways integrate clusters and supercomputers from all over the world.
  - IU_PTI can make your resources available to your team.
  - IU_PTI can help you access supercomputers, clusters, and computing clouds from outside your institution or enterprise.

This slide courtesy V4I.us; may be used
Only if attributed as:
Slide courtesy of the Virtual Verification
Validation and Visualization Institute.
V4I.US

WASHINGTON BIG A TECHNOLOGY CONFERENCE

Advanced Assurance in Manufacturing

Virtual Verification Validation & Visualization Institute
v4i.us

V4I is managed by the
National Center for Defense Manufacturing and Machining (NCDMM)

V4I

- Mission: Enable the use of digital data, modeling and simulation across supply chains to accelerate the introduction of new materials, manufacturing processes and product systems & services to market while meeting demanding regulatory requirements.
- V4I.us
- A private – public partnership brought together to collaboratively solve problems common to private sector partners

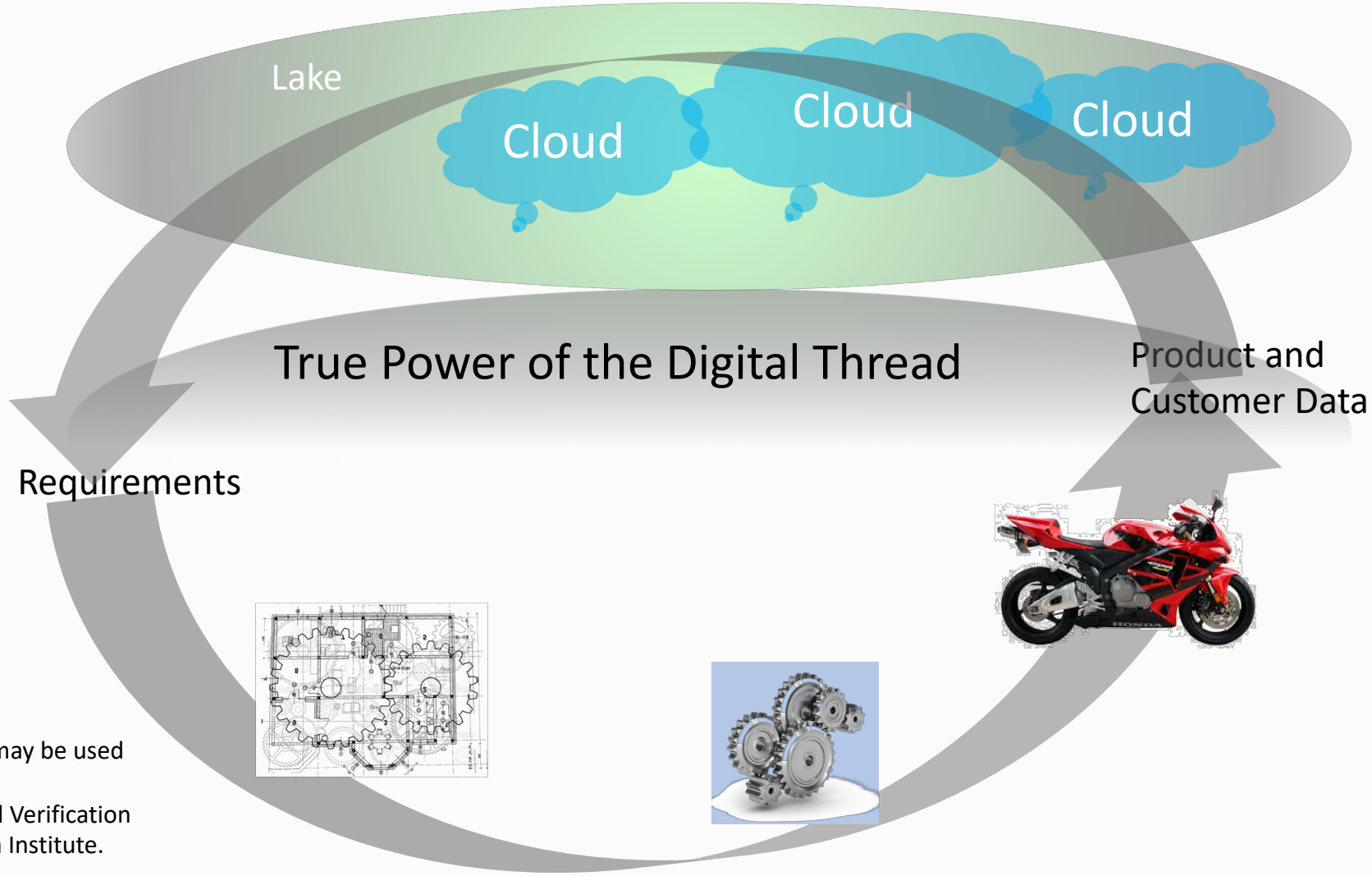
This slide courtesy V4I.us; may be used only if attributed as:
Slide courtesy of the Virtual Verification Validation and Visualization Institute. V4I.US

V4I Context and Rationale

- Virtual: Existing outside of (for example: digitally, in graphic or computational form) and representing a physical reality.
- Verification – “The evaluation of whether or not a product, service, system or model thereof complies with a regulation, requirement, specification, or imposed condition.”
- Validation – “The assurance that a product, service, system or model thereof meets the needs of the customer and other identified stakeholders.”
- Visualization – “The formation of mentally accessible images; the act or process of interpreting in visual terms or of putting into visible form.”

- V4I Value Proposition: Increasing the scientific use, reliability, and effectiveness of virtual testing reduces cost and time to market (for the benefit of private sector members, who can use V4I tools to increase competitiveness without enabling its supply chain to aid competitors)

Research: Big Data & Virtual Customer



Industry Value – Sample ROI values for investment in digital product verification

- Defense Aerospace
  - 50% Research and Development cost savings
  - 25% Research and Development time reduction
- Life Sciences: Medical Devices
  - 50% Research and Development cost savings
  - 50% Research and Development time reduction
- Improved Product Quality, Safety, Reliability
  - Higher customer Satisfaction
  - Sustainable Product Lines and Lifecycles through Innovation
- Improved Manufacturing Process Safety, Reliability, Efficiency
  - Higher return on investment

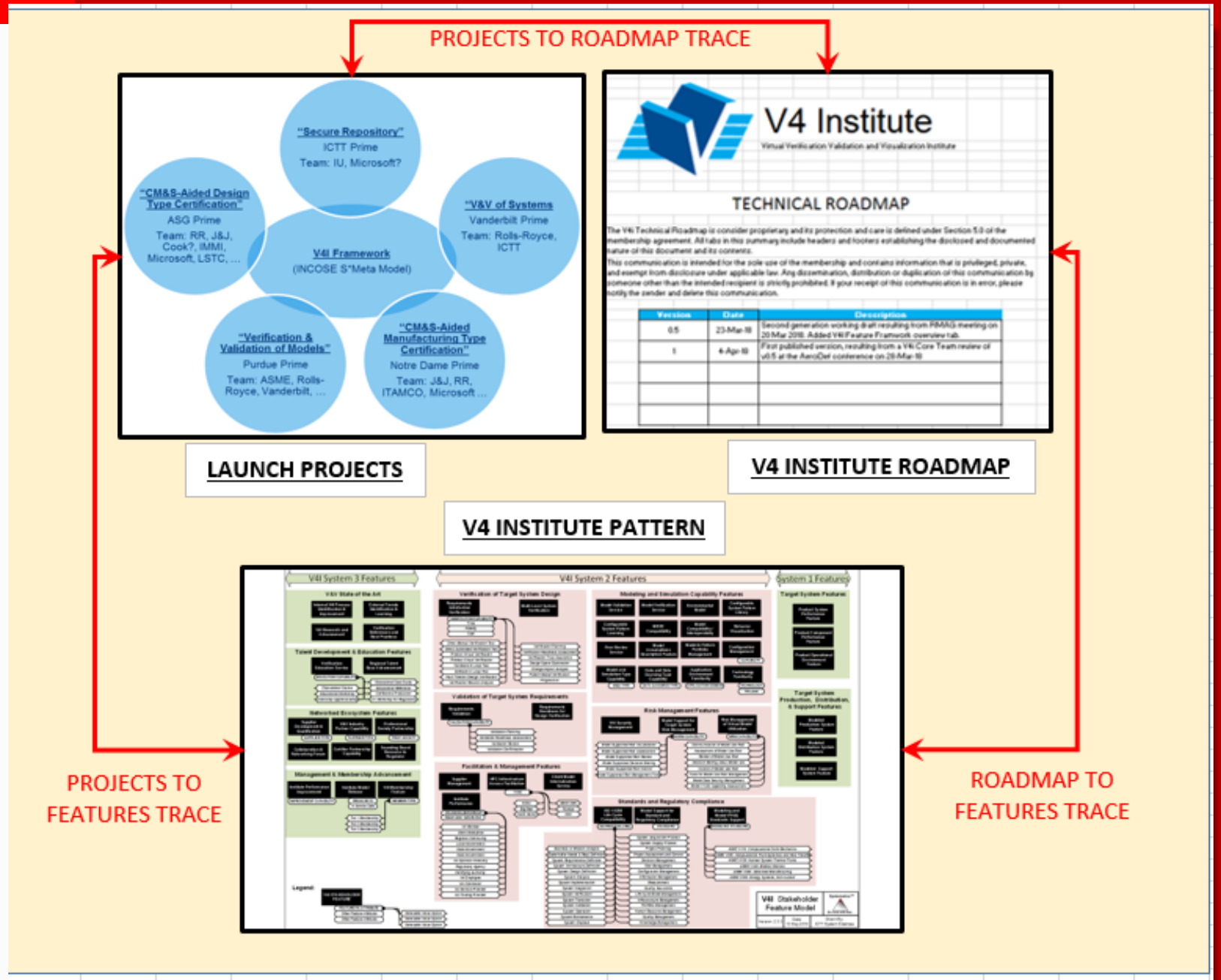
Why do this as a collaboration

- Option: Isolated Private Corporations Only?
  - Each company arrives at the same scientific solution independently
- Option: University or Government Research Only
  - Application of science to business practical solutions an issue
- Option: Entrepreneurial Network, or Specialized Solution Providers
  - Costs high, competitive IP conflicts, limited horizon
- But one has to make this safe to share in the supply chain without competitive harm

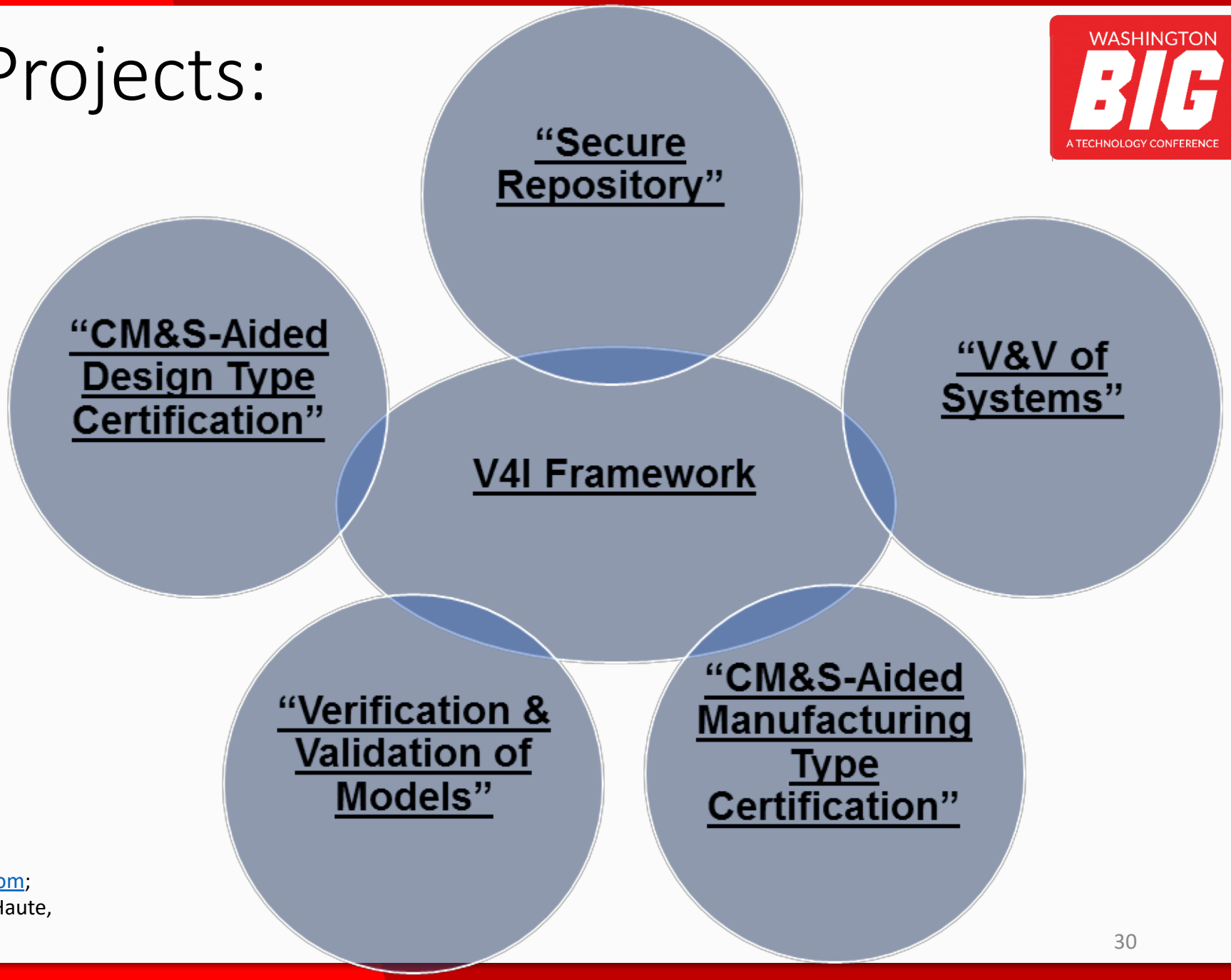
Regional Value for the Midwest – Sustainable Economic Impact

- Manufacturing Excellence – Return on Investment, Realized Innovation
- Safety, Quality, Reliability, Cost – Better products, processes, safer jobs
- Education Realization – Research opportunities, STEM durability
- Regulatory Efficiency – Clear decision: impact to public safety & confidence
- Entrepreneurial Networks – Opportunities, Agility, Markets, Job creation
- Community – Stability, Continuity

V4I Systems Engineering Processes



V4I Launch Projects:



Features of Model Repository

Persistent Model Repository

Repository features providing persistent (memory) services for virtual models of systems

Model Credibility Support

Repository features supporting credibility of shared models

Compatibility

Repository features supporting the compatibility of virtual models across representations, media, and life cycle tooling

Community of Interest and Teaming Support

Repository features supporting collaboration and community use of models

Fit to User and Use

Repository features associated with fitness of the Repository to intended use and users

Model Application Transaction Support

Repository features supporting repository transactions specific to targeted applications of the repository

Model-Based Pattern Support

Repository features supporting reusable configurable model-based patterns

Model Life Cycle Management

Repository features supporting management of the life cycle of models.

Repository System Sustainability

Repository features sustaining its sustainability

Configurable Repository Deployment

Repository features supporting varied deployment options

Repository Release Roadmap

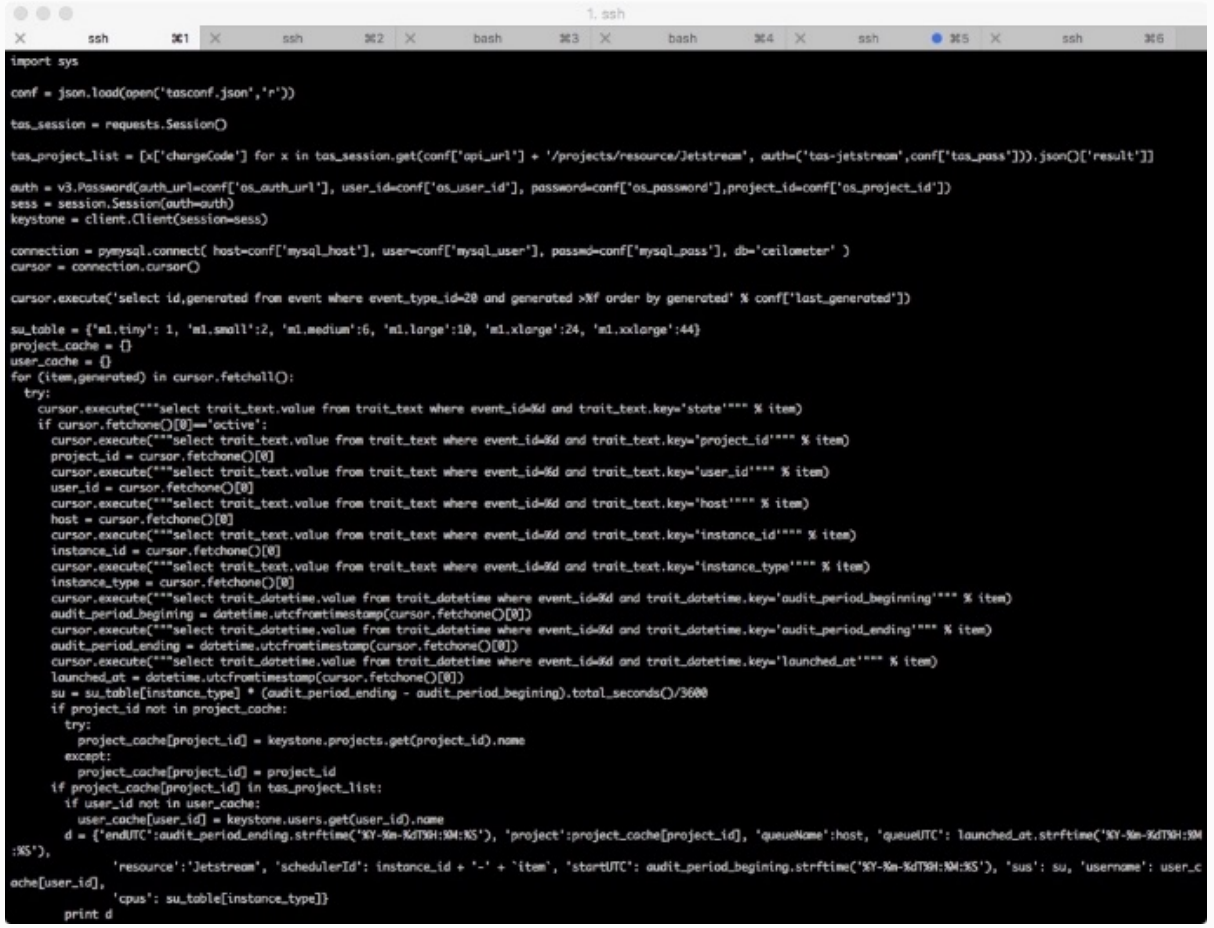
Repository features release plan

Notes:
- This pattern is a specialization of the S*Repository Pattern.
- A specific repository configuration may have a populated subset of the features of this pattern framework.
- This is a Feature pattern describing the trade space for such repositories, but the appearance of a Feature in the above model is no guarantee that it can be supported in an implemented system, unless it appears in the configured Features for that system.

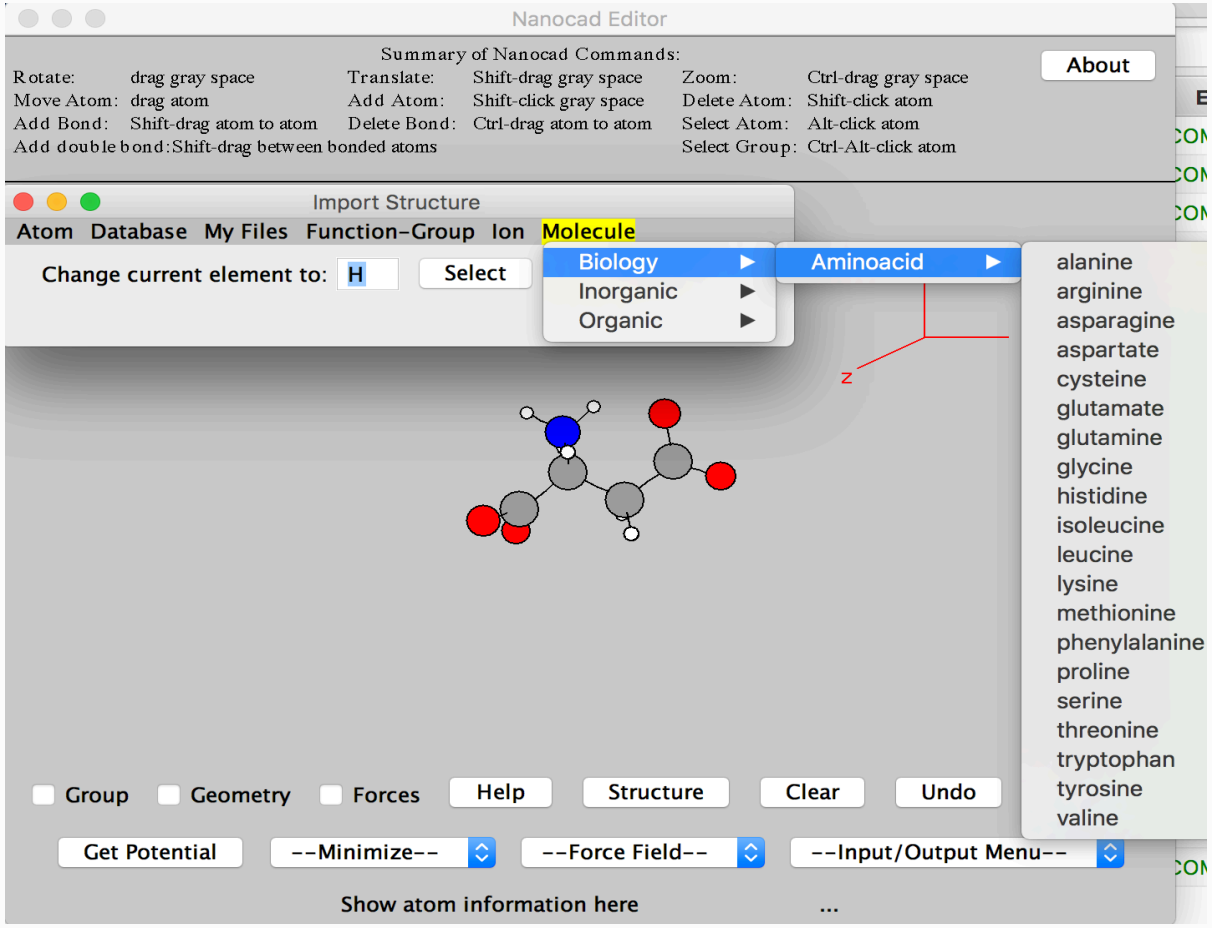
V4I Model Repository Feature Model: Summary Feature Groups		Systematica™ Do more with less
Version: 1.2.1	Date: 30 Oct 2017	Drawn By: ICTT System Sciences

This slide courtesy
William Schindel,
May be reused
Only if attributed as:
Slide courtesy of the William Schindel;
schindel@ictt.com;
ICTT System Sciences,
378 South Airport Street,
Terre Haute, IN 47803.

Technology Adoption Choices

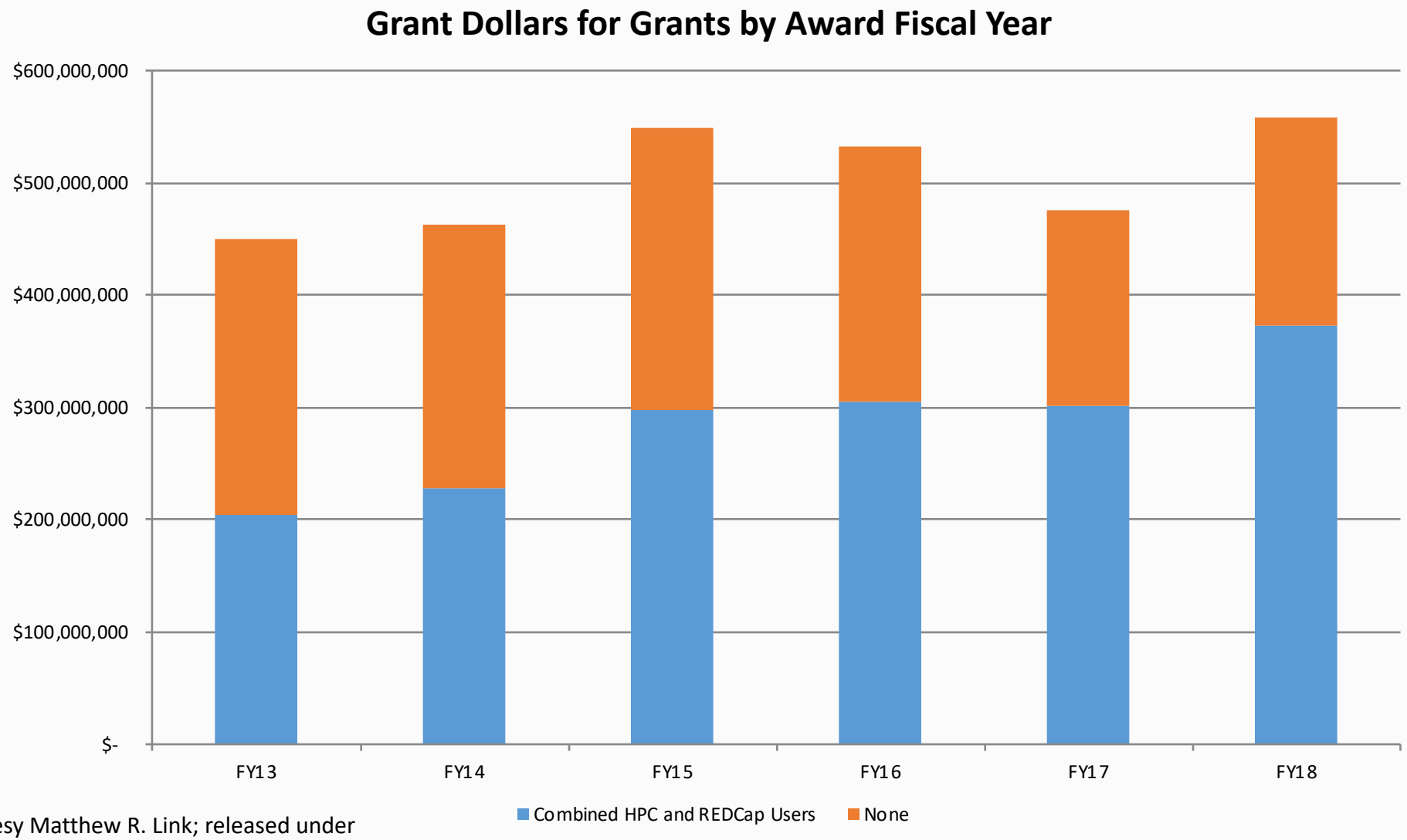


```
import sys

conf = json.load(open('tasconf.json','r'))

tas_session = requests.Session()

tas_project_list = [x['chargeCode'] for x in tas_session.get(conf['api_url'] + '/projects/resource/Jetstream', auth=('tas-jetstream',conf['tas_pass'])).json()['result']]

auth = v3.Password(auth_url=conf['os_auth_url'], user_id=conf['os_user_id'], password=conf['os_password'],project_id=conf['os_project_id'])
sess = session.Session(auth=auth)
keystone = client.Client(session=sess)

connection = pymysql.connect( host=conf['mysql_host'], user=conf['mysql_user'], passwd=conf['mysql_pass'], db='ceilometer' )
cursor = connection.cursor()

cursor.execute('select id,generated from event where event_type_id=20 and generated >%f order by generated' % conf['last_generated'])

su_table = {'m1.tiny': 1, 'm1.small':2, 'm1.medium':6, 'm1.large':10, 'm1.xlarge':24, 'm1.xxlarge':44}
project_cache = {}
user_cache = {}
for (item,generated) in cursor.fetchall():
  try:
    cursor.execute("""select trait_text.value from trait_text where event_id=%d and trait_text.key='state'""" % item)
    if cursor.fetchone()[0]=='active':
      cursor.execute("""select trait_text.value from trait_text where event_id=%d and trait_text.key='project_id'""" % item)
      project_id = cursor.fetchone()[0]
      cursor.execute("""select trait_text.value from trait_text where event_id=%d and trait_text.key='user_id'""" % item)
      user_id = cursor.fetchone()[0]
      cursor.execute("""select trait_text.value from trait_text where event_id=%d and trait_text.key='host'""" % item)
      host = cursor.fetchone()[0]
      cursor.execute("""select trait_text.value from trait_text where event_id=%d and trait_text.key='instance_id'""" % item)
      instance_id = cursor.fetchone()[0]
      cursor.execute("""select trait_text.value from trait_text where event_id=%d and trait_text.key='instance_type'""" % item)
      instance_type = cursor.fetchone()[0]
      cursor.execute("""select trait_datetime.value from trait_datetime where event_id=%d and trait_datetime.key='audit_period_beginning'""" % item)
      audit_period_begining = datetime.utcfromtimestamp(cursor.fetchone()[0])
      cursor.execute("""select trait_datetime.value from trait_datetime where event_id=%d and trait_datetime.key='audit_period_ending'""" % item)
      audit_period_ending = datetime.utcfromtimestamp(cursor.fetchone()[0])
      cursor.execute("""select trait_datetime.value from trait_datetime where event_id=%d and trait_datetime.key='launched_at'""" % item)
      launched_at = datetime.utcfromtimestamp(cursor.fetchone()[0])
      su = su_table[instance_type] * (audit_period_ending - audit_period_begining).total_seconds()/3600
      if project_id not in project_cache:
        try:
          project_cache[project_id] = keystone.projects.get(project_id).name
        except:
          project_cache[project_id] = project_id
      if project_cache[project_id] in tas_project_list:
        if user_id not in user_cache:
          user_cache[user_id] = keystone.users.get(user_id).name
        d = {'endUTC':audit_period_ending.strftime('%Y-%m-%dT%H:%M:%S'), 'project':project_cache[project_id], 'queueName':host, 'queueUTC': launched_at.strftime('%Y-%m-%dT%H:%M
:%S'),
             'resource':'Jetstream', 'schedulerId': instance_id + '-' + `item`, 'startUTC': audit_period_begining.strftime('%Y-%m-%dT%H:%M:%S'), 'sus': su, 'username': user_c
ache[user_id],
             'cpus': su_table[instance_type]}
        print d
```



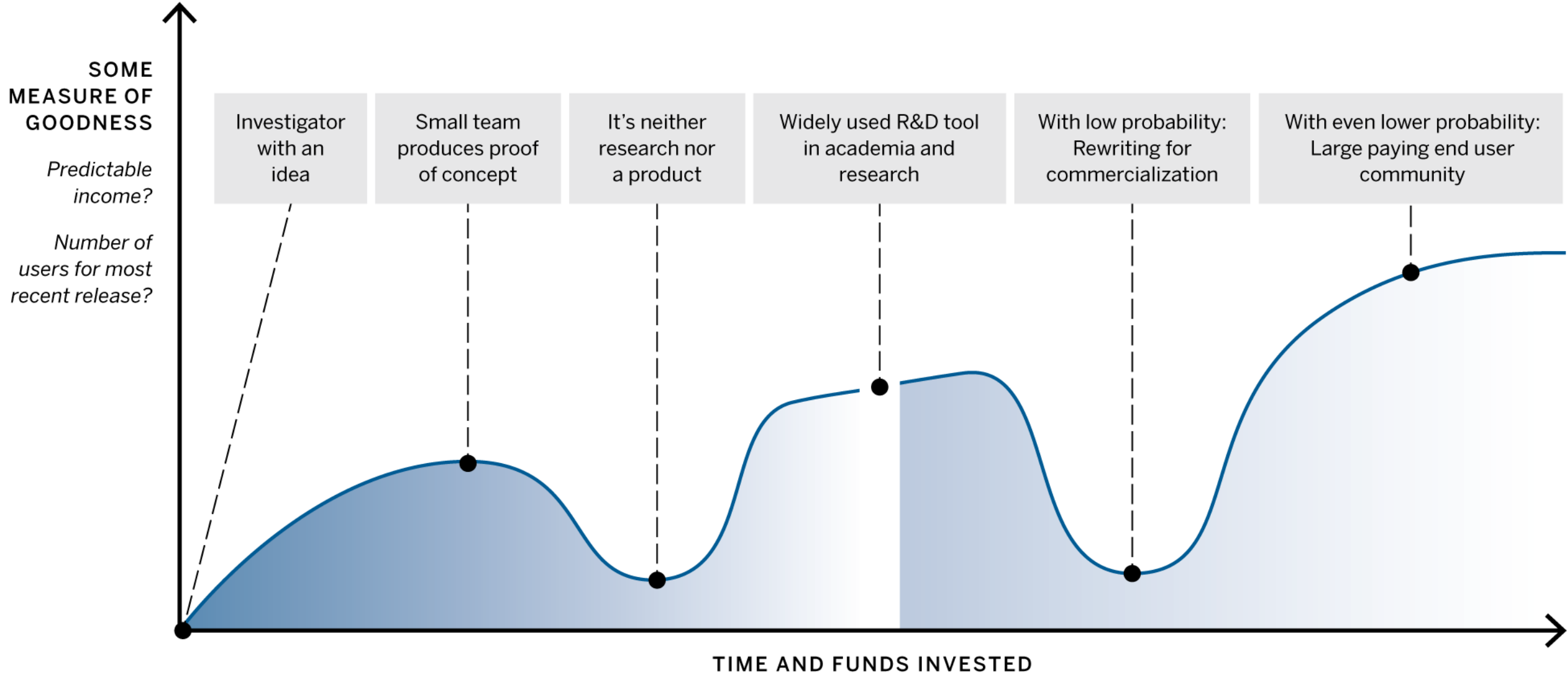
This slide courtesy Sudhakar Pamidighantam, pamidigs@iu.edu; released under default license for this presentation.

Does this stuff all make fiscal sense for IU?

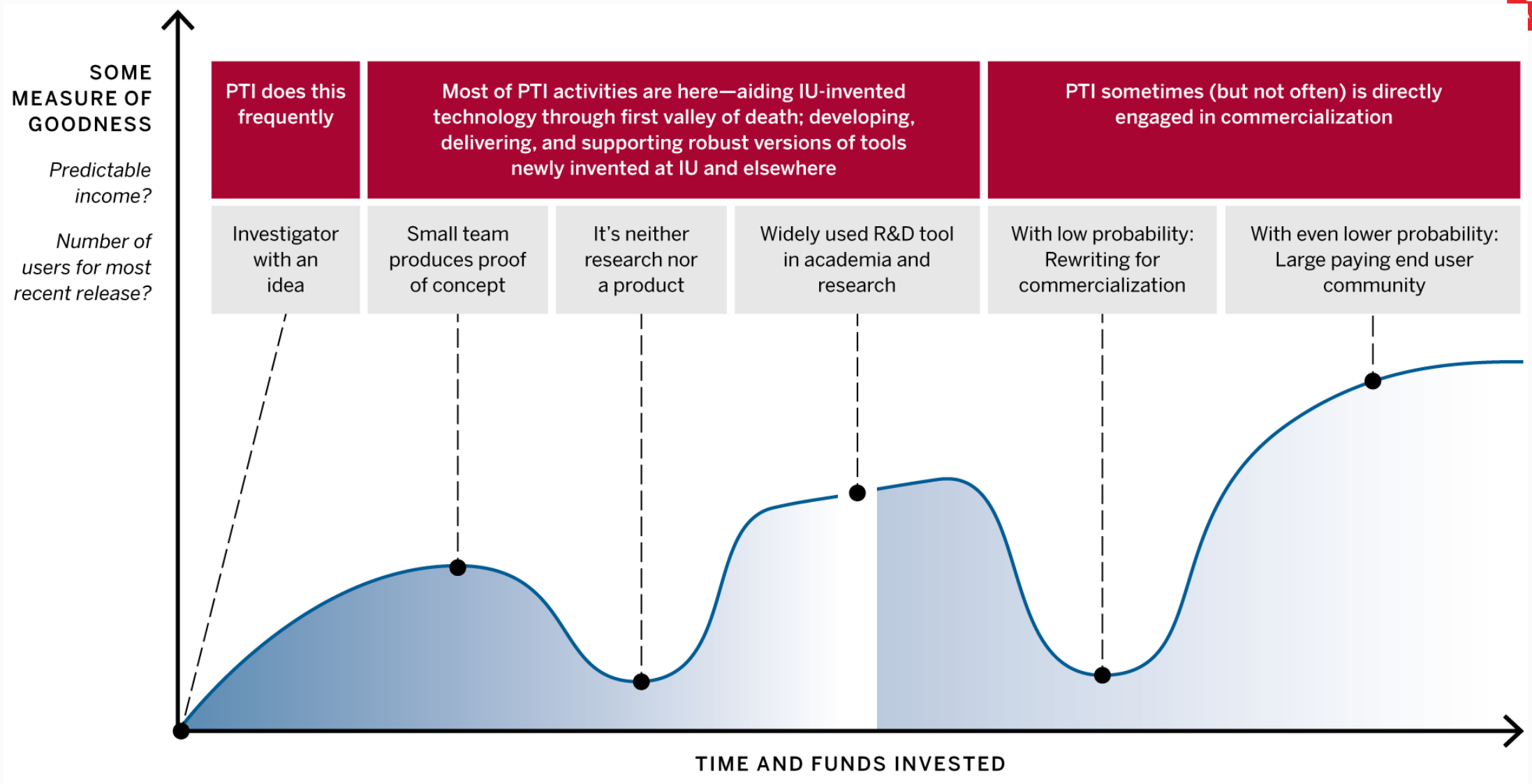


This slide courtesy Matthew R. Link; released under default license for this presentation.

The *two* valleys of death



PTI takes services and tools through the two valleys of death

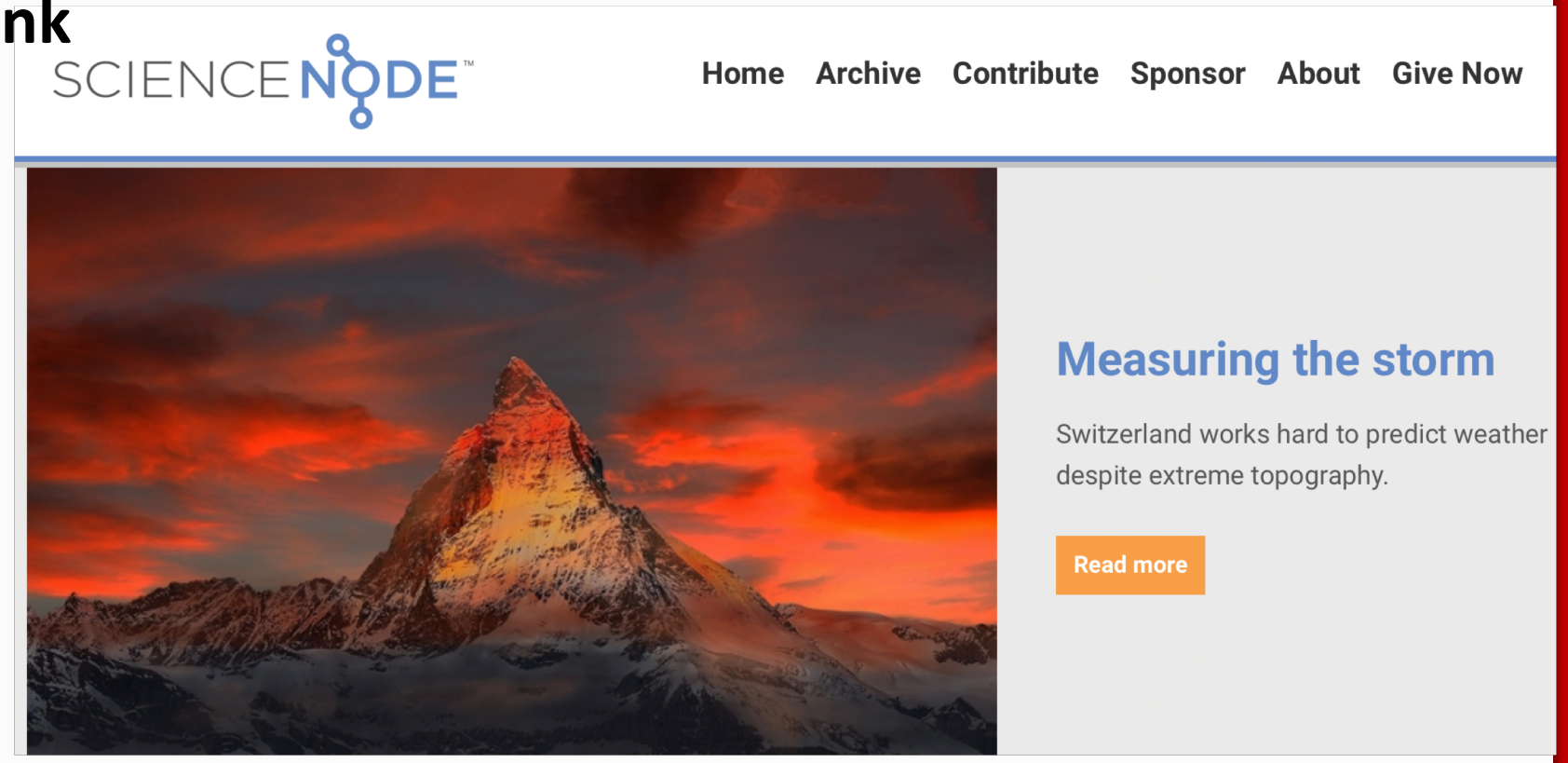


Acknowledgments

- The Indiana University Pervasive Technology Institute was created in 1999 by a major gift from the Lilly Endowment and persists today through a combination of competitively obtained federal funding, donations, and IU support
- XSEDE (the eXtreme Science and Education Discovery Environment) is supported by NSF award 1053575 (John Towns, UIUC, PI); XSEDE in turn supports some of the Science Gateway research described here.
- Jetstream is supported by NSF award 1445604 (David Y. Hancock, IU, PI; Craig Stewart, founding PI)
- Science gateway development is supported by a number of grant awards from the NSF, most particularly Award 1339774 (Marlon Pierce, PI)
- Several slides used today are derivatives of slides created by V4I
- Opinions presented here are those of the author(s) and do not necessarily represent the views of the NSF, IUPTI, IU, or the Lilly Endowment, Inc.

Utterly Shameless Plugs

- Follow PTI on Twitter at @iu_pti **(I will post a link to these slides on Twitter!)**
- Follow @scinode on Twitter
- Sciencenode.org

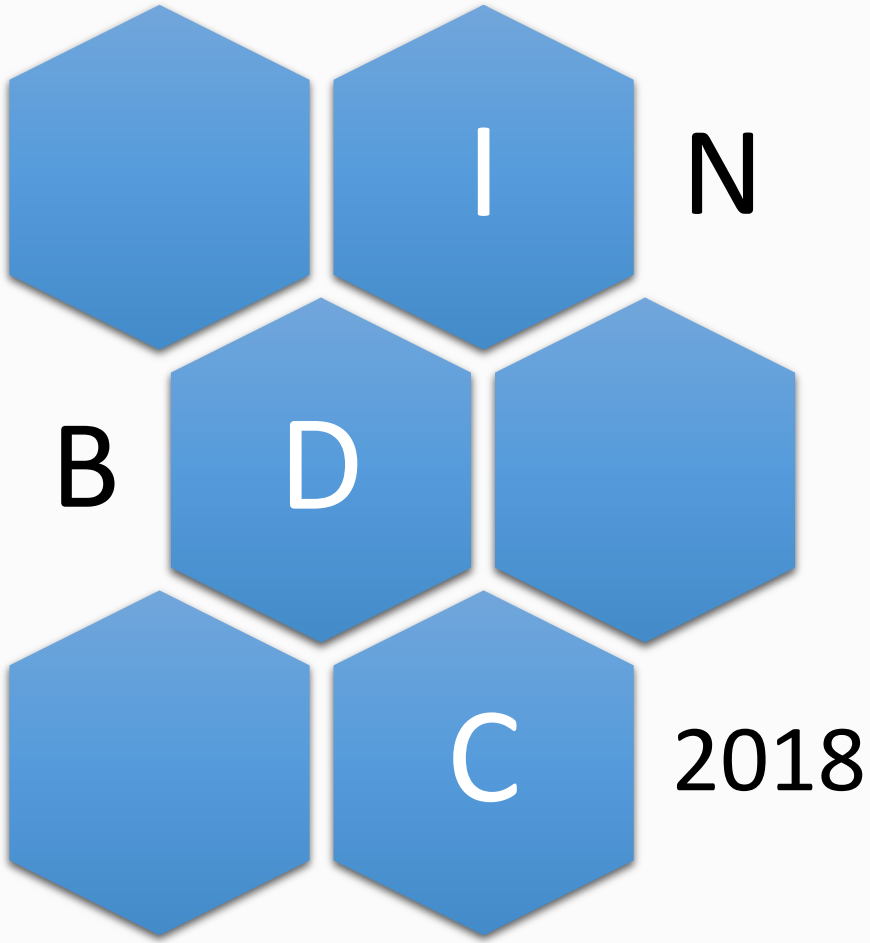


INDY BIG DATA
A TECHNOLOGY CONFERENCE

Questions ?

INDY BIG DATA

A Technology Conference



INDY BIG DATA
A TECHNOLOGY CONFERENCE

THANK YOU